Issue 75 September 1997

Metadata

by Rachel Heery, Andy Powell **and** Michael Day
Metadata Projects Group, UKOLN The UK Office for Library and Information Networking, University of Bath

'Metadata' has become a fashionable and overused term, but nevertheless provides a useful label within the library world for description of digital resources. It is an important part of the activity being undertaken to impose some order on the explosion of material available across networks. This Briefing examines metadata within the context of network information management and describes some of the growing number of projects and services which are now using metadata for resource discovery in a networked environment.

INTRODUCTION

Why metadata?

Some people do not like the term 'metadata'. Metadata means subtly different things within the various disciplines that use the term. It has also become a fashionable term, and is often overused. We would argue, however, that it is a label useful within the library world for referring to information about resources, and in particular description of digital resources. There is a different emphasis within the computer science disciplines, where the term refers to data which describe data elements, datasets or database management systems, and where metadata models and metadata systems are constructed to integrate disparate databases. One can see overlaps between such work and resource discovery and information management, but there are marked differences in the nature of the data described: the unit being described would be a data element in computer science, and a resource in the information world. In the information world metadata may consist of an agreed set of data elements with agreed semantics, agreed syntax and agreed rules for formulating the content of the elements.

The term metadata is useful in that it acknowledges a significant change in the emphasis between traditional book cataloguing and the activity being undertaken today to impose some order on the explosion of material available across networks. Caplan points out the advantages of using a 'new' term that does not have the traditional connotations of cataloguing.[1] The popularity of its usage is indicative of the interest in resource description running across both computer science and librarianship. It reflects changes in the nature of cataloguing brought about by digital technology, changes which David Levy typifies as 'cataloguing in the digital order'.[2]

Within this Briefing we will be examining metadata in the context of what is traditionally called bibliographic control but might more widely be understood as network information management. We will use metadata to mean the information about a resource which enables us to identify, locate and request that resource. Metadata also allows us to manage resources, both in terms of local database management, and access management (for example controlling terms and conditions of access). Metadata can be 'descriptive data', such as author, title; 'subject data', such as uncontrolled keywords or controlled language descriptors; 'access data', describing hardware and software requirements for using a resource; and metadata might also be 'administrative data', describing the metadata itself, such as who created the record, date the record was created, owner of the metadata record. It might also include information about terms and conditions of use. The range of metadata as described here illustrates that metadata is itself 'data' and, particularly in the context of system design, is not usefully distinguished from other data.[3]

What is metadata for?

Much activity is centred on development of metadata formats and the standardization of these formats. The emphasis on formats should not obscure the importance of the process requirements—metadata cannot be viewed in isolation from the context in which it is used. Within information systems metadata performs a range of functions. These include :

- *Searching*: identifying the existence of a resource by keyword searching, browsing indexes or visualization techniques.
- *Location*: finding a particular instance of a re source.
- *Selection*: analysis and evaluation based on the description provided.
- *Semantic interoperability*: allowing searching across domains by means of equivalent elements.
- *Resource management*: collection and database management.
- *Terms of availability information*.

Decisions about formats will be influenced by which of the above functions the metadata will perform. Thus within a system it will sometimes be appropriate to have a simple metadata format, for example to allow

for interoperability in searching across subject domains, while on occasion a richer format will be required to enable selection of resources in a specialized domain.

Much of this Briefing will concentrate on metadata as it relates to networked resources and in particular to World Wide Web resources. The opportunities provided by the Web for new services and new publishing processes require new forms of resource description. The volatile nature of Web documents and the continuing increase in the amount of information being made available are driving services to seek alternatives to high cost traditional cataloguing. Services looking at incorporating the advantages of an automated approach to indexing are tending towards the use of simple resource description formats.

Resource discovery service models

In the context of the Web, users are offered alternative options for discovering resources, all of which are based more or less on structured metadata. These include:

- *Lists*: lists of pointers to useful resources.
- *Searching*: by keyword or controlled vocabulary.
- *Browsing*: alphabetically by subject keyword, or using more formal subject classification schemes.
- *Visualization*: navigation of the Web site by spatial browsing techniques.

At present the predominant service for discovery of Web resources is the *search engine* or *search service* which may use one or more of these techniques. Search engines can be categorized by their coverage and selection policy, and by the method by which their indexes are created. A number of search services have been evaluated although the lack of information on policies available from the larger services make comparisons difficult.[4,5]

Coverage of search engines can be characterized as:

- *Global*: these would attempt to cover all Web sites, although in reality this will be limited in terms of granularity and frequency of update.
- *Geographical*: covering all Web sites in a particular area, country or region.
- *Sectoral*: this might be a subject area, a user community like higher education, or a curatorial tradition like museums, libraries or archives.
- *Selective*: typically sectoral services will select resources for description on the basis of quality criteria.
- *Organizational/Intranet*: organizations or individuals may want to allow searching of their own resources.

The indexes on which these services are based may be derived from the automatically harvested full text of Web resources, or they may be based on records created manually. Pilot implementations are now beginning to make use of metadata embedded in resources, in particular Dublin Core embedded in HTML. In the future it seems likely that more metadata will be held on Web sites independently of the HTML, or on third party databases linked to the Web resource.

Range of formats

When examining the issues surrounding the use of metadata within the Web environment, it is helpful to consider the wider context of resource discovery. Metadata formats vary according to a number of criteria and there is increasing awareness of the strengths and weaknesses of these various diverse formats. Metadata ranges from generic simple Internet resource descriptions to highly structured records relating to complex objects such as databases. On the one hand there is the full text indexing of the global search services (Excite, Lycos, etc.) where the complete text of Web documents is indexed, there is no fielded record, and the 'display record' is an extract from the full text, typically the first few lines. On the other hand there are the complex tagged record of MARC formats, or the analytical mark-up of SGML-based formats.

Detailed reviews of current metadata formats have been carried out elsewhere.[6,7] Here we will present a

simple typology of formats along a continuum from simple to rich (Figure 1).

Depending on the position on this continuum from simple to rich it is possible to associate a number of characteristics with the three bands of metadata and these are summarized briefly in Figure 2. The simplest formats are used to create relatively unstructured indexes for locating items, whereas the most complex records can be used as the basis of sophisticated analysis and navigational tools. The simpler records are created automatically and the more complex by hand. This will affect the overall cost of record creation. Simpler records do not permit complex designation of sub-fields and qualifiers whereas the richer records have defined rules for detailed designation of sub-fields. The more complex formats are associated with relatively heavy-weight search and retrieve protocols (like Z39.50), whereas the simpler formats tend to be associated with directory service protocols.

DUBLIN CORE

Dublin Core history

The Dublin Core Element Set (Dublin Core or DC) is a fifteen element metadata set that is primarily intended to aid resource discovery on the Web.[8] Dublin Core forms a simple description record, which has emerged as a result of a series of workshops sponsored by the Online Computer Library Center (OCLC) and other organizations:

- OCLC/NCSA Metadata Workshop, Dublin, Ohio. March 1995.
- OCLC/UKOLN Warwick Metadata Workshop, Warwick. April 1996.
- CNI/OCLC Image Metadata Workshop, Dublin, Ohio. September 1996.
- Fourth Dublin Core Workshop, Canberra. March 1997.

FIGURE 1: A SIMPLE TYPOLOGY OF RESOURCE DISCOVERY METADATA

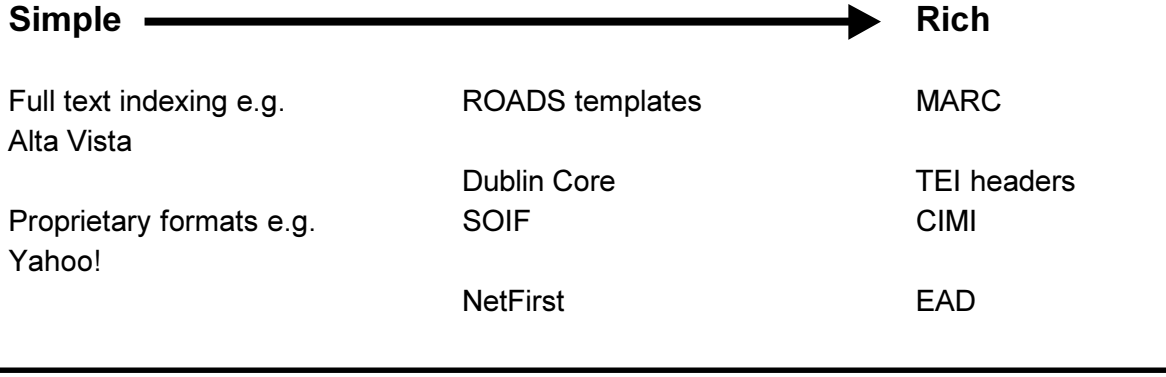
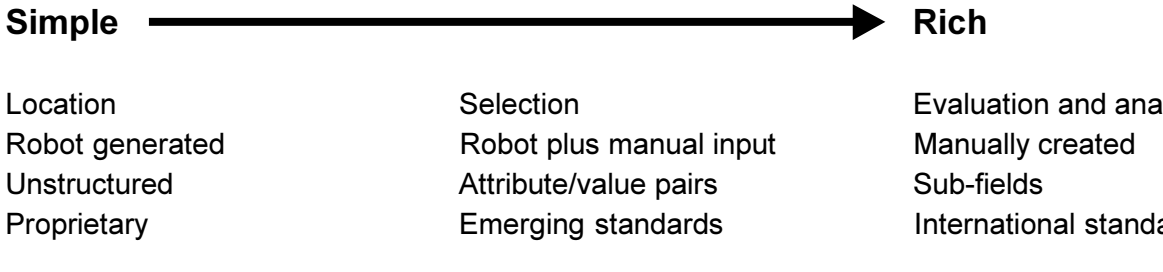


FIGURE 2: ASSOCIATED CHARACTERISTICS OF METADATA FORMATS



The workshops represent a consensus building effort which has included participants from a range of backgrounds (IETF, SGML, digital library research), domains (text, image, geographic information systems) and professions (librarians, computer scientists, content specialists). This consensus and the international acceptance of Dublin Core are probably the most significant outcomes of the workshops, and have largely been achieved through the leadership of OCLC.

The objectives for Dublin Core set by the first workshop were firstly to define a simple set of data elements so that authors and publishers of Internet documents could create their own metadata with no extensive training—the Dublin Core approach being mid-way between the detailed tagging of MARC or structured TEI headers and the automatic indexing of locator services such as Alta Vista. Secondly, Dublin Core aimed to provide a basis for semantic interoperability between other, more complicated, formats. By means of mapping from more complex formats, and by 'filtering' more complex formats, Dublin core facilitates searching across other disparate record formats.

An initial element set was agreed upon and certain principles were established for further development of the set, these being :

- *Extensibility*: the core set can be extended with further elements as it is acknowledged that many 'publishers' or metadata producers may wish to augment this simple set with more specialized data.
- *Optionality*: all elements are optional.
- *Repeatability*: all elements are repeatable.

During the first workshop there was an explicit decision not to define syntax at that stage, but first to reach consensus on the semantics of a minimum element set. To tie Dublin Core semantics to any one particular syntax (as in the MARC family of record formats) was seen as unhelpful. The second workshop, which took place in the UK at the University of Warwick in April 1996 sponsored by UKOLN and OCLC, went on to consider possible syntaxes. Embedding metadata in resources using HTML was the obvious choice to fulfil the immediate need of pilot implementations.

The Warwick workshop also looked at the implementation of Dublin Core and requirements for extensibility, change control and implementation. The Warwick framework emerged as a concept from the second workshop. This is a model for a container architecture for packages of metadata, each package being metadata of a different type.[9,10]

The third workshop, the CNI/OCLC Image metadata workshop, considered use of the Dublin Core element set for describing images, in particular those images which could be defined as 'document like objects'. Perhaps surprisingly the workshop reached the conclusion that images could be described using the minimal Dublin Core elements with some minor adjustments.

Discussion prior to the fourth workshop in Canberra resulted in agreement to extend the thirteen elements agreed in Dublin to fifteen, and these fifteen have been defined and documented in an Internet-Draft.[11] Note that all Dublin Core elements are optional, so you do not have to embed all fifteen elements into each Web page. They can also be repeated if necessary, for example to indicate that a page has more than one author.

The semantics of Dublin Core elements can be modified using qualifiers, and use of qualifiers was central to discussions at the Canberra workshop.[12] There are three kinds of qualifier: the TYPE qualifier which refines the meaning of an element; the SCHEME qualifier which indicates that the element value conforms to some external and widely recognized scheme; and the LANGUAGE qualifier which indicates the language of the element value. It has been agreed that the use of qualifiers should refine the element rather than extend it. In general, the intention is that a Web robot should be able to take the embedded Dublin Core metadata, throw away all of the qualifiers and still have something meaningful to add to its index. However, the widespread use of qualifiers could cause severe problems with interoperability.

The marked confidence in Dublin Core has had significant impact on standards-making activities such as USMARC discussions, Z39.50, and W3C initiatives;

it has also been chosen as the solution for early implementations within projects in Australia, Scandinavia, Europe and the US.

Dublin Core creation and management

By embedding Dublin Core metadata into Web pages and then gathering it into searchable databases using Web robots it will be possible to provide Web-based search services with improved precision over those currently available.

In order for Web page authors and Web-site administrators to be able to embed Dublin Core metadata into Web pages there need to be tools available.[13] As an aid to creating Dublin Core META tags several Web based 'Dublin Core generators' have been made available on the Web. One of these is DC-dot, available from the UKOLN Web-site.[14] DC-dot first prompts for the URL of the Web page that you want to describe. It then retrieves that page from the Web and automatically generates Dublin Core META tags to describe it. The Dublin Core META tags are then displayed in such a way that they can be updated and extended manually using a Web form. Once editing is complete the tags can be copied into a Web page using cut-and-paste to a text editor. Alternatively, DC-dot will convert the Dublin Core into other formats, including USMARC, SOIF, XML, IAFA/ROADS, and send these formats back to you via your Web browser or e-mail.

However, the last few years have seen a general move away from using simple text editors to create and maintain HTML pages towards the use of more sophisticated authoring tools. These tools do not, in general, make it easy to add META tags to Web pages. Even where tools do allow for the creation of META tags there are longer term issues associated with embedding metadata by hand that must be considered. What happens if the syntax for embedding metadata in HTML changes in the future? How easy will it be to move embedded metadata into alternative metadata formats that are likely to become more commonly used in the future, for example in PICS-NG?

More recently Web-site management tools have become available which hold all the pages for a site in a database. A 'publish' button causes the information in the database to be written out as a set of HTML Web pages. These tools have the immediate advantage of standardizing the style of Web pages across a site, and in future may become metadata aware. In the meantime the use of these tools for managing metadata may be possible using available 'macro' facilities.

Sites interested in home grown solutions to the issues of managing metadata may choose to hold the metadata separately, in a neutral format, and then convert it and embed it into Web pages using 'server-side include' scripts. A more detailed description about one such system being implemented at UKOLN is available elsewhere.[15]

WEB INDEXES

Harvesting

Once Dublin Core metadata is embedded into significant numbers of HTML Web pages it needs to be collected into a Web index so that it can be made available using a *search engine*. This may be done on a site-wide basis, to form a local site search engine, or it may be done across a group of Web servers to form a more comprehensive search engine encompassing, for example, all the Web pages in a geographical region or subject area. The collection of metadata from Web pages is usually done using a Web robot. A Web robot can be thought of as an automated Web browser. Starting from a given URL or set of URLs it visits each page in turn extracting the embedded metadata and adding it into a database (Web index). For each page visited, the robot also extracts all the embedded links in the page and adds them into a list of URLs still to be visited. The robot needs to maintain this list of URLs in such a way that it does not visit the same server too often in quick succession, thus overloading it, but also needs to ensure that pages are revisited fairly regularly so that information in the

METADATA IN HTML

HTML allows arbitrary metadata to be embedded into the *head* section using the META tag. To make things clearer, here is an example:

```
<HTML>
<HEAD>
<TITLE>UKOLN: UK Office for Library and Information Networking</TITLE>
<META NAME="Keywords" CONTENT="national centre, network information support, library
community, awareness, research, information services, public library networking,
bibliographic management, distributed library systems, metadata, resource discovery,
conferences, lectures, workshops">
<META NAME="Description" CONTENT="UKOLN is a national centre for support in
network information management in the library and information communities. It provides
awareness, research and information services and is based at the University of Bath">
</HEAD>
<BODY>
...
</BODY>
</HTML>
```

In this example, the TITLE tag and the two META tags give the title, some keywords and a short description for the page. Note that the HTML specification does not say anything about what type of metadata should be placed into the META tags. However, the Web robots used by some of the big Internet search engines (for example Alta Vista) look for the two META tags shown in this example and use them to improve the effectiveness of their searches. Words found in theses tags are given extra weight when they match user queries and pages with these tags tend to appear higher up in search results than pages without them. Because of this, these two META tags are in fairly common usage.

DUBLIN CORE IN HTML

The elements in the Dublin Core are TITLE, SUBJECT, DESCRIPTION, CREATOR, PUBLISHER, CONTRIBUTOR, DATE, TYPE, FORMAT, IDENTIFIER, SOURCE, LANGUAGE, RELATION, COVERAGE and RIGHTS. These elements can be embedded into META tags in the *head* section of a Web page in a similar way as the example above. Here is the same page with embedded Dublin Core tags:

```
<HTML>
<HEAD>
<TITLE>UKOLN: UK Office for Library and Information Networking</TITLE>
<META NAME="DC.title" CONTENT="UKOLN: UK Office for Library and Information Networking">
<META NAME="DC.subject" CONTENT="national centre, network information support, library
community, awareness, research, information services, public library networking,
bibliographic management, distributed library systems, metadata, resource discovery,
conferences, lectures, workshops">
<META NAME="DC.description" CONTENT="UKOLN is a national centre for support in network
information management in the library and information communities. It provides
awareness, research and information services and is based at the University of Bath">
<META NAME="DC.creator" CONTENT="UKOLN Information Services Group">
</HEAD>
<BODY>
...
</BODY>
</HTML>
```

database does not become out of date. For large search engines covering many Web sites it may be necessary to run several Web robots on several machines, all feeding metadata into the same database, in order to increase the rate at which Web pages can be indexed.

This is exactly how the big search engines, like Alta Vista, function. However, their Web robots do not currently look for embedded Dublin Core and thus have to extract the available metadata in the form of Keywords and Description META tags or try to automatically generate metadata based on the text of the HTML page or simply build a full-text index. In many cases a combination of these three approaches is taken.

In the case of building a search engine for a single Web-site it may not be necessary to run a Web robot to collect metadata. The Web index can be built directly from the files on the Web server filestore. This is the approach taken by the public domain CNIDR Isite software.[16] Isite is an integrated Internet publishing software package including a text indexer, a search engine and Z39.50 communication tools to access databases.[17] It is worth noting that there are a couple of problems in building an index based directly on files rather than by using a Web robot. Firstly, a filestore view of a Web server may include many pages that are not visible on the Web (because they are not linked to any other pages). It may well be undesirable to include such pages in a Web index. Secondly, metadata that is embedded using server side includes (SSI) will not be available to a program that simply reads a file from the Web filestore.

Although none of the big search engines looks for embedded Dublin Core metadata, there are some projects that are developing robots that do. The European DESIRE project is building a partial European Web index, covering the Nordic countries, using a Web robot that is being enhanced to extract embedded Dublin Core metadata.[18] Similarly, the UK Electronic Libraries Programme (eLib) NewsAgent for Libraries project will obtain information content for the service by the use of a Web robot that will look for embedded Dublin Core and other—NewsAgent—specific metadata in pages.[19] The eLib ROADS project, which provides the tools used by the other eLib 'subject services' to construct databases of Internet resource descriptions, will also use this software to construct robot-generated ROADS databases. There are other projects around the world looking at similar areas.[20] Some of these projects are described in more detail later in this Briefing.

Distributed searching

Having collected metadata using a Web robot, it needs to be made available for searching. There are several approaches to this. A fundamental concept is that of *centralized* verses *distributed* searching. A centralized search engine pulls all the metadata into a single database. Although this database may be mirrored in several places, users only have the opportunity of searching one database at a time. Alta Vista is an example of a centralized Web index. A distributed Web index is made up of a group of databases that may well be physically distributed across the Internet. In addition to sharing the load across multiple servers this approach also allows for localized management of server databases. Searches may be sent in parallel to all the databases and the results merged, or may be routed to appropriate databases in some way.

There are various protocols available to facilitate distributed searching, including Z39.50, WHOIS++, and LDAP (described below). These protocols enable a client to send a search request to a server and obtain results from several databases. Depending on the protocol and the contents of the underlying database, the client may be able to request more detailed information about the search results (which may initially be returned as a simple list of hits) and may also be able to request that the full text of the object be returned. In some cases the client may be a dedicated piece of software, for example a Java applet or a Web browser plug-in, running on the end user's local computer. Often, however, the search client will be a CGI based gateway running on a Web server and accessed by the end user as a Web based form.[21,22]

The DESIRE European Web Index, following the

distributed model, is made available using several GILS compliant Z39.50 servers, one per country. Users indicate which of the servers they would like their search sent to as part of specifying the search. Results from multiple servers are merged before being displayed to the user.

In the ROADS project, distributed ROADS databases are made available using the WHOIS++ protocol. Searches across several ROADS databases (both robot-generated and manually constructed) are possible with searches currently being sent to each server in parallel. Future versions of the ROADS software will support the Common Indexing Protocol, which allows servers to share knowledge about their databases, and thus route queries between different servers in a more efficient manner.[23] It should be noted that the Common Indexing Protocol is not specific to WHOIS++ and could be used, in theory, to route queries between multiple LDAP servers or multiple Z39.50 servers.

PROJECTS AND SERVICES USING METADATA

There are a growing number of projects and services currently using metadata for resource discovery in a networked environment. The following section comprises a brief description of some of these projects.

Projects funded by the Electronic Libraries Programme

Access to Network Resources projects

The UK Electronic Libraries programme (eLib), a series of projects, demonstrators and services funded by the Joint Information Systems Committee (JISC) of the UK Higher Education Funding Councils, was formed in 1995 in response to recommendations made by the authors of the Report of the Joint Funding Councils' Libraries Review Group in December 1993—the Follett Report. Amongst other things, the Report recommended that JISC should fund the 'development of a limited number of top level networking navigation tools in the UK to encourage the growth of local subject based tools and information servers'.[24] Once eLib was in place, it funded several Access to Network Resources (ANR) projects and services.[25] These include:

- ADAM: Art, Design, Architecture & Media Information Gateway;
- Biz/ed: Business Education on the Internet;
- EEVL: Edinburgh Engineering Virtual Library;
- IHR-Info: Institute of Historical Research;
- OMNI: Organizing Medical Networked Information;
- RUDI: Resources for Urban Design Information;
- SOSIG: Social Science Information Gateway.

These projects are creating large amounts of metadata for network resources in their specialist areas. These subject services, sometimes called subject-based information gateways, are one solution to the problem of resource discovery on the Internet. The services use specialist staff to select Internet resources ensuring quality control, and these are then described using human-created metadata. The subject service approach to resource discovery is based to some extent on the traditional library model. Resources are chosen according to defined selection criteria and they will then be manually 'catalogued' for inclusion in a database. This process ensures that only good quality resources are made available through the service and that sufficient metadata is available to enable the adequate searching and retrieval of these resources. The resulting service often provides access both by searching and by browsing, either by a list of subject terms or by a particular subject-classification. Several of the eLib subject services are based on the software tools developed by the ROADS project.

ROADS: Resource Organization and Discovery in Subject-based services

ROADS is an eLib project, also under the ANR strand, and is a collaboration between the Institute of Learning and Research Technology (ILRT) at the University of Bristol, the UK Office for Library and Information

Networking (UKOLN) at the University of Bath and the Department of Computer Studies at Loughborough University.[26] Its aim is to develop and implement a user-orientated resource discovery system enabling users to find and access networked resources. In short, ROADS is developing discovery software for a networked discovery framework primarily with regard to the requirements of the eLib ANR services.

ROADS is very much concerned with metadata—its creation, organization and also how it can be searched and presented to users. ROADS templates, the metadata format chosen for use by the ROADS project, are based on IAFA/WHOIS++ templates—a format originally designed for anonymous FTP archives. They are based on simple (text based and human readable) attribute/value pairs of variable length. One major advantage of using ROADS templates is the possibility of searching across multiple subject services using the WHOIS++ protocol.[27]

The nature of the ROADS project has resulted in its participation in wider discussions of metadata and Internet resource discovery. For this reason, ROADS partners have been involved with the Dublin Core initiative and with deployment of WHOIS++. There is also a strong focus on the semantic interoperability of metadata formats: producing metadata mappings or crosswalks, looking at potential interaction with the Z39.50 protocol; the development of template registries, cataloguing rules, etc.

NewsAgent

NewsAgent for Libraries is another eLib project, this time in the Electronic Journals programme area.[28] The aim of the project is to create a user-configurable electronic news and current awareness service for library and information professionals—the information content being taken from selected UK library and information science journals and briefing materials from five organizations. The service will obtain information content from a Web robot designed to look for embedded Dublin Core and other—NewsAgent specific—metadata. As part of the project, UKOLN have developed a replacement for the HTML *summarizer* that is available as part of the Harvest suite of resource discovery tools. This work is intended to make the Harvest Web robot Dublin Core aware and will eventually be made available with the public domain version of the Harvest software.[29]

European Union funded projects

DESIRE: Development of a European Service for Information on Research and Education

The DESIRE Project is an extremely large project funded by the EU Telematics for Research Sector of the Fourth Framework Programme.[30] The project is investigating Web technology and the implementation of pilot information services on behalf of European researchers and is divided into ten work packages. The one with the most relevance to metadata issues is work package 3 (WP3), 'Resource discovery and indexing',[31] which has the general aim of supporting research users of the Internet to locate information relevant to their research. The work package partners include all of the ROADS project partners, together with NetLab (University of Lund, Sweden) and the National Library of the Netherlands. It has two main strands:

- *Subject services (subject-based information gateways).* Building on the subject service approach to Internet subject services in conjunction with work done at NetLab on engineering (EELS—Engineering Electronic Library, Sweden) and the National Library of the Netherlands(NBW—Nederlandse Basisclassificatie Web), WP3 has looked at quality-controlled subject-based information gateways based on library-type selection and cataloguing skills. A demonstrator is planned for European social science information, together with further services for engineering and fine art.
- *Automated indexing of WWW information sources.* WP3's work on providing tools and methods for the automatic indexing of the WWW information is an extension of work carried out at NetLab and the National Technological Library of Denmark (DTV) on the Nordic Web Index (NWI). A European Web Index (EWI) will be developed as part of WP3 to provide a harvesting and indexing

service for the academic sector in Europe and to establish a single uniform service with the aim of indexing all European Internet documents relevant to the academic area.

Several reports have been produced as part of the project. NetLab have produced a state-of-the-art review of indexing and data collection methods used in robot-based Internet search services[32] and a functional specification for a European Web Index.[33] WP3 has also resulted in a three-part report on a *Specification for resource description methods* which included a survey of current metadata formats, a study of quality selection criteria for Internet subject services, and an evaluation of the use of subject classification schemes for providing access to Internet resources.[34]

BIBLINK: Linking Publishers and National Bibliographic Services

The BIBLINK project is funded by the Telematics Applications Programme of the European Commission and aims to create an electronic link between publishers of electronic material and national bibliographic agencies.[35] The project is led by the British Library, and its partners include the national libraries of France, The Netherlands, Norway and Spain, the Universitat Oberta de Catalunya in Barcelona, and UKOLN. The intention of the project is that the bibliographic experts of the national libraries of Europe, with cooperation of partners in the book industry, will be able to examine what type of descriptive metadata would be required for catalogues of electronic publications and to investigate the possibility of establishing electronic links for the transfer of this metadata from publishers to national bibliographic agencies. BIBLINK intends to produce an interactive demonstration system which would enable selected electronic publishers to transmit metadata to national bibliographic agencies, where this data would then be enriched and converted to specific MARC formats (primarily UNIMARC and UKMARC) for use by national libraries. The level of data required is the minimum amount sufficient to support traditional Cataloguing in Publication (CIP) type functions.

There are two distinct phases in BIBLINK, the second one involving the development and installation of the demonstration system at the sites of the project partners and participating publishers. The first phase, however, consists of a series of seven work packages investigating background issues for BIBLINK. Work package 1, for example, made recommendations regarding what particular formats should be accepted from publishers, deciding to look at SGML DTDs like Simplified SGML for Serial Headers (SSSH) for complex records and the use of Dublin Core as a minimum element set for data exchange.[36] Work Package 2 reviewed the important area of unique identifiers for electronic publications, including the Uniform Resource Name (URN), the Serial Item and Contribution Identifier (SICI) and the Digital Object Identifier (DOI).[37] Other work packages have looked at the transmission of data between libraries and publishers, conversion processes to investigate interoperability between publishers' metadata and MARC formats, and the important area of authentication.

Other metadata-related projects and services

Nordic Metadata Project

The Nordic Metadata Project is funded by NORDINFO, the Nordic Council for Scientific Information, and has six participating organizations.[38] The Nordic countries are used to sharing information about printed materials, but there is an awareness that sharing information about electronic documents has been complicated by the inadequacy of current resource discovery mechanisms. The project is using Dublin Core, and amongst other things, is investigating the following:

- The production of conversion tables and programs to convert Dublin Core to Nordic MARC formats. An experimental converter can currently produce NORMARC, FINMARC and USMARC records. Other Nordic formats will be added to the converter, together with a MARC to DC converter, if required. It is intended that the software should also be able to be easily adapted to convert DC to non-Nordic MARC formats.

- The production of tools for the creation of Dublin Core metadata to encourage an improvement in the quality and quantity of metadata that is made available. A Nordic Metadata DC production template was published at the start of 1997 and has since been modified to conform with the changes to HTML syntax agreed at the DC 4 Workshop in Canberra.
- Working with the DESIRE project to make the Nordic Web Index robot metadata aware so that it can recognize and extract embedded Dublin Core.

The range of activities being carried out by the Nordic Metadata Project—metadata creation, harvesting and interoperability—will be of great interest to others who are considering the implementation of metadata-based systems.

Arts and Humanities Data Service

The Arts and Humanities Data Service (AHDS) is funded by JISC for the collection, description and preservation of the electronic resources that result from and are used by research and teaching in the humanities.[39] It consists of an executive based at King's College London, and five service providers, located throughout the UK:

- Archaeology Data Service (A consortium, led by the University of York);
- History Data Service (The Data Archive, University of Essex);
- Oxford Text Archive (Oxford University Computing Services);
- Performing Arts Data Service (Glasgow University);
- Visual Arts Data Service (Surrey Institute of Art and Design).

AHDS will provide a unified catalogue giving access to its service provider's holdings and possibly to other scholarly collections. For this reason, the AHDS has examined the needs of arts and humanities scholars with regard to information discovery and resource description with the intention of identifying shared metadata requirements which could be used in a distributed catalogue.[40] AHDS, in conjunction with UKOLN, initiated Resource Discovery Workshops in early 1997 so that specific requirements in all relevant disciplines could be integrated into a system giving access to a distributed, interdisciplinary and mixed-media collection of digital resources.[41] It is recognized that each service provider may have its own preferred formats for storing metadata; for example, the Oxford Text Archive will be using TEI headers. The AHDS is looking at a solution where a core set of metadata, based on Dublin Core, could be used to provide 'top-level' access to the distributed AHDS resource, while individual service providers maintain their own specific metadata for their own collections. It is possible that the subject-specific metadata created by service providers could be used to generate automatically (through metadata mappings/crosswalks) a subset of core metadata which could then be used in a 'top-level' catalogue.

The MathN Broker

A service currently using metadata is the MathN Broker—a mathematical pre-print service based at the University of Osnabrück, Germany.[42] The service grew out of a '*Fachinformation*' project run by the DMV, the German Mathematical Society. The service gives electronic access to PostScript versions of pre-prints stored on about 40 departmental Web servers in Germany.[43] The Harvest software is used for indexing, but this has limitations when used with PostScript. For this reason, the pre-print service indexes metadata which was originally stored in what Roland Schwänzl has described as a 'preliminary Warwick Container for HTML coded MetaData', using a format known as the MathDMV-Preprint Core.[44] Since the beginning of 1997 the service has used Dublin Core elements embedded in HTML META tags. The metadata can include subject classifications from the Mathematics Subject Classification (MSC), the Physics and Astronomy Classification Scheme (PACS) and the ACM Computing Classification System (CCS), together with subject keywords and abstracts. The metadata is provided by authors using a Web page with a FORMS interface called the Mathematics Metadata Markup editor (MMM).

ASSOCIATED TECHNOLOGIES

Protocols

HTTP

The Hypertext Transfer Protocol (HTTP) defines the way in which Web clients (typically Web browsers such as Netscape Navigator) and Web servers communicate with each other. It specifies how clients request a particular page from a server—such requests are based on Uniform Resource Locators (URLs). It enables clients to ask for information about a page, such as when it was last updated. It also specifies how servers send Web pages, informational messages and error messages back to the client.

LDAP

The Lightweight Directory Access Protocol (LDAP) was developed as a simple alternative to the ISO X.500 protocol, a protocol for providing distributed information about people—names, e-mail addresses, telephone numbers, etc. Although primarily designed for providing access to information about people, LDAP can also be used for other sorts of information—for example, to access data about Web pages. LDAP servers are typically organized into a strict hierarchy with the 'root' at the top, country level nodes below that, organizational nodes below them, etc.

WHOIS++

The WHOIS++ protocol was developed as a lightweight Internet protocol for providing distributed information about people—names, e-mail addresses, telephone numbers, etc. It can also be used for other sorts of information. The eLib ROADS project provides software that uses WHOIS++ to distribute descriptions of Internet resources. Unlike LDAP and X.500, WHOIS++ does not have a strict hierarchical representation of the data space, instead using a more flexible 'mesh' of servers. WHOIS++ based searches are routed through this mesh based on 'forward knowledge' held by one server about another. This 'forward knowledge' is maintained using the Common Indexing Protocol (CIP).

Z39.50

Z39.50 is a standard for information retrieval approved by the National Information Standards Organization (NISO), a committee accredited by the American National Standards Institute (ANSI). It has also been recognized by the International Organization for Standardization (ISO) where it is known as ISO 23950. Z39.50 can be described as a protocol for supporting the construction of distributed information retrieval applications.[45] The protocol allows client applications (known in the standard as the 'origin') to search databases on remote servers (the 'target') and to retrieve relevant information. As an open standard, Z39.50 supports the retrieval of information from distributed remote databases.[46] The first applications were developed specifically for bibliographic data, for example the distributed searching of library online public access catalogues, but attribute-sets can be defined to allow the protocol to work with many other types of data.

Languages

HTML

The HyperText Markup Language (HTML) is the language in which World Wide Web documents are written and is an application of the Standard Generalized Markup Language (SGML).[47] HTML is primarily concerned with two things: defining how documents look—by the use of a variety of structural or presentational tags; and the creation of hypertext links to separate network documents. HTML pages are split into two main sections, the header or HEAD element and the BODY. The HEAD section of a page contains information about the document (or metadata), for example an HTML TITLE tag, while the BODY will typically contain the information content of the document itself together with its structural and presentational tags—which can then be displayed by a Web browser.

XML

XML stands for 'Extensible Markup Language', and is a simplified subset of SGML. Development of XML is an initiative within W3C (the World Wide Web Consortium) and its aim is to define an SGML DTD for the Web.[48] XML is designed to allow flexibility and extensibility (hence the name). Whereas HTML facilitates display of information on the Web, XML provides for standards-based management of data (including metadata). XML specifies how the semantics of data elements can be expressed, indicating what each data element means. Examples of XML tags might include author, price, person lastname, person firstname and so on, there being no limit on the tags that might be included in a schema. XML is a text-based markup language similar to HTML to look at, but indicating the semantics of data rather than specifying mode of display.

Schema specifying agreed element names can be shared between Web 'publishers', and the schema can itself be expressed in XML. XML can be used to add semantic information to an HTML document, and HTML using devices such as stylesheets, can display the information expressed in XML in a standardized way. XML looks likely to be used within various applications now under development for publishing information on the Web: for the Meta-Content Framework (MCF), the Channel Definition Format (CDF) and the new version of PICS labels.

PICS (Platform for Internet Content Selection)

Another initiative within W3C is PICS which currently provides a mechanism for associating numeric content rating labels with Internet resources.[49] PICS enables attributes to be linked to a resource and rated on a numeric scale (e.g. level of violence = 10). PICS is now being used primarily as a means to filter content on the Web particularly against criteria such as suitability for children. The next version of PICS (commonly referred to as PICS-NG) will provide an infrastructure for associating more general string labels (i.e. metadata) with resources. It is likely that XML will be used as the language for encoding PICS labels.

PICS does not define semantics but is positioned as a transport syntax (i.e. a syntax for sharing data between applications). It is envisaged that different element sets might be encoded using PICS-NG, and that Dublin Core might be one of these. PICS records might be embedded in the resource, linked to the resource, or indeed located independently on a third party database.

IDENTIFIERS

Unique identifiers are an essential part of the technology that enables electronic trading, copyright management, electronic tables of contents, production tracking and resource discovery. Traditionally publishers and libraries have worked with identifiers such as the ISBN and ISSN for paper products. These identifiers are assigned at the book or journal level, but the need for a unique and persistent identifier for electronic resources at a lower level of granularity has become more important. Increasingly, we need to identify much smaller fragments of complete works, for example parts of text, images, video clips, pieces of software, etc. Recent schemes, such as the DOI, can be used at arbitrary levels of granularity determined by individual publishers based on commercial or other considerations.

There are significant outstanding issues in relation to identifiers:

- What is being identified? For an online document that has multiple versions and that is mirrored on several Web sites, is it the logical 'document' that is being identified or particular instances of that document?
- Identification vs. location. The Uniform Resource Locator (URL) that we are all familiar with is a locator rather than an identifier. If an object moves, its associated URL changes and people using the

old URL are likely to get a failure indicating that it is no longer available. There are significant political and commercial interests which act as barriers to establishing services which will resolve identifiers to URLs.

ISSN (International Standard Serial Number)

The ISSN is a standardized international numeric code which enables the identification of serial publications, for example periodicals, newspapers, annuals or series. Serials can be in printed form, on other medium (microform, floppy disk, CD-ROM or CD-i), or can be accessible online. An ISSN is normally represented as the string 'ISSN' followed by two sets of four digits: for example, **ISSN 0374-0536.**

ISBN (International Standard Book Number)

The ISBN system is an international standard numbering system for monographs. It has traditionally been used for books, but has been expanded to include other new media such as videocassettes and electronic media. An ISBN is normally represented as the string 'ISBN' followed by ten digits separated into four parts: for example, **ISBN 82-7111-124-8.**

SICI (Serial Item and Contribution Identifier)

The SICI is a variable length code that uniquely identifies serial issues (items) and articles within a serial (contributions). The SICI is a complex identifier split into three parts: the item segment (based on the ISSN of the serial); the contribution segment (which identifies an article or other contribution within the serial); and the control segment. For example: **0730-9295(199206)11:2<168:CRFAOC>2.0.TX;2-#.**

PII (Publisher Item Identifier)

Elsevier Science developed the PII to identify journal articles independently from their packaging unit, because they may be published in different ways (database, CD-ROM, paper, World Wide Web, etc.). It is primarily intended for document items of interest to scientific publishers. For example: **S0165-3806(96)00403-8.**

URN (Uniform Resource Name)

Uniform Resource Names (URNs) are intended to serve as persistent, globally unique resource identifiers that fit into the larger Internet information architecture composed of, additionally, Uniform Resource Characteristics (URCs) and Uniform Resource Locators (URLs). URNs are for identification, URCs for including metadata and URLs for locating resources. URNs are designed to make it easy to map other identification schemes into URN-space. The exact format of URNs is still under discussion but it is likely that, for example, an ISBN may be represented as a URN as follows: **urn:isbn:0-395-36341-1.**

DOI (Digital Object Identifier)

The Digital Object Identifier (DOI) system is being developed on behalf of the Association of American Publishers (AAP). The DOI system is based around a directory, which stores an object's DOI and its associated location (URL). Queries sent to the directory result in the DOI being looked up and the location returned to the client. In Web terminology, this is a standard Hypertext Transfer Protocol (HTTP) redirect. A DOI has two parts, a globally unique part called the Publisher ID and a publisher assigned part called the Item ID. For example: **10.153/34571.**

PURL (Persistent Uniform Resource Locator)

PURLs have been developed and deployed by OCLC as a naming and resolution service for general Internet resources. Functionally, a PURL is an URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL Resolution Service associates the PURL with the actual URL and returns that URL to the client. The client can then complete the URL transaction in the normal fashion. As with the DOI this is achieved using an HTTP redirect. For example: **http://purl.oclc.org/OCLC/PURL/INET96.**

GLOSSARY

CIMI Computer Interchange of Museum Information. CIMI records are an SGML-based metadata format developed for museum information.

DTD Document Type Definition. An application program defining document types in an SGML context.

Dublin Core Dublin Core Metadata Element Set. A metadata format defined on the basis of international consensus which has defined a minimal information resource description, generally for use in a Web environment.

EAD Encoding Archival Description. An SGML-based metadata format developed for the description of archives.

GILS Government Information Locator Service. Metadata format created by the US Federal Government in order to provide a means of locating information generated by government agencies.

Granularity The level of detail at which indexing takes place.

Harvest A system providing a set of software tools for the gathering, indexing and accessing of Internet information. Uses SOIF.

IAFA templates Internet Anonymous FTP Archive templates. Metadata format designed for anonymous FTP archives, now adapted for use in ROADS project.

MARC MAchine Readable Cataloguing. A family of formats based on ISO 2709 for the exchange of bibliographic and other related information in machine readable form.

PICS Platform Independent Content Selection. Internet content filtering infrastructure.The next generation (PICS-NG) is likely to provide a general metadata infrastructure.

ROADS Resource Organization and Discovery in Subject-based services. eLib funded project developing software for use by Internet subject services.

SGML Standard Generalized Markup Language. An international standard (ISO 8879) for the description of marked-up electronic text.

SOIF Summary Object Interchange Format. A metadata format developed for use with the Harvest architecture.

SSI Server Side Includes. A mechanism for dynamically generating parts of Web pages.

TEI Text Encoding Initiative. An attempt to define, using SGML, the encoding of literary and linguistic texts in electronic form. TEI headers are an SGML-based metadata format used for the documentation of these texts.

Warwick Framework An architecture for the exchange of distinct metadata packages involving the aggregation of metadata packages into containers.

REFERENCES

1. See: Caplan, P. 'You call it corn, we call it syntax-independent metadata for document-like objects'. *The Public-Access Computer Systems Review*, 6(4), 1995. Available from: <URL:http://info.lib.uh.edu/pr/v6/n4/capl6n4.html>

2. Levy, D. *Cataloging in the digital order*. [Paper for: Digital Libraries '95: The Second Annual Conference on the Theory and Practice of Digital Libraries, Austin, Texas, June 11-13, 1995]. Available from: <URL:http://csdl.tamu.edu/DL95/papers/levy/levy.html>

3. Lagoze, C. 'From static to dynamic surrogates: resource discovery in the digital age'. *D-Lib Magazine*, June 1997. Available from: <URL:http://www.dlib.org/dlib/june97/06lagoze.html>

4. Stobart, S. and Kerridge, S. 'An investigation into World Wide Web Search Engine use from within the UK—preliminary findings'. *Ariadne*, 6, November 1996. Available from: <URL:http://www.ariadne.ac.uk/issue6/survey/>

5. Koch, T., Ardö, A., Brümmer, A. and Lundberg, S., *The building and maintenance of robot based Internet search services: a review of current indexing and data collection methods*. Draft D3.11 (version 3) for Work Package 3 of Telematics for Research project DESIRE, September 1996. Available from: <URL:http://www.ub2.lu.se/desire/radar/reports/D3.11/>

6. Dempsey, L. and Heery, R., with contributions from M. Hamilton, D. Hiom, J. Knight, T. Koch, M. Peereboom and A. Powell, *Specification for resource description methods—Part 1: A review of metadata: a survey of current resource description formats*. Deliverable 3.2 (1) for Work Package 3 of Telematics for Research project DESIRE, March 1997. Available from: <URL:http://www.ukoln.ac.uk/metadata/DESIRE/overview/>

7. Burnard, L. and Light, R., *Three SGML metadata formats: TEI, EAD, and CIMI*. A study for BIBLINK Work Package 1, December 1996. Available from:<URL:http://www.ukoln.ac.uk/metadata/BIBLINK/wp1/sgml/>

8. *The Dublin Core Metadata Element Set: Home Page*. Available from: <URL:http://purl.org/metadata/dublin_core>

9. Lagoze, C., Lynch, C. and Daniel, R., *The Warwick Frame work: a container architecture for aggregating sets of metadata*. TR96-1593, June 21, 1996. Available from: <URL:http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR96-1593>

10. Lagoze, C. 'The Warwick Framework: a container architecture for diverse sets of metadata'. *D-Lib Magazine*, July/August 1996. Available from: <URL:http://www.dlib.org/dlib/july96/lagoze/07lagoze.html >

11. Weibel, S., Kunze, J. and Lagoze, C., *Dublin Core Metadata for simple resource description*. Internet-Draft, 9 February 1997. Available from: <URL:ftp://ds.internic.net.internet-drafts/draft-kunze-dc-00.txt>

12. Weibel, S., Iannella, R. and Cathro, W., 'The 4th Dublin Core Metadata Workshop Report: DC-4, March 3—5, 1997, National Library of Australia, Canberra'. *D-Lib Magazine*, June 1997. Available from: <URL:http://www.dlib.org/dlib/june97/metadata/06weibel.html>

13. *UKOLN Metadata Software Tools*. Available from: <URL:http://www.ukoln.ac.uk/metadata/software-tools />

14. *DC-dot*. Available from: <URL:http://www.ukoln.ac.uk/metadata/dcdot/>

15. Powell, A., 'Dublin Core management'. *Ariadne*, 10, July 1997. Available from: <URL:http://www.ariadne.ac.uk/issue10/dublin/>

16. *CNIDR Isite*. Available from: <URL:http://vinca.cnidr.org/software/Isite/Isite.html>

17. *BSn Doctypes Description.* Available from: <URL:http://w3.bsn.com/Z39.50/INTRO.html>

18. *Nordic Web Index.* Available from: <URL:http://nwi.ub2.lu.se/>

19. Powell, A., *Notes on use of Dublin Core by NewsAgent*. Available from: <URL:http://www.ukoln.ac.uk/metadata/NewsAgent/dcusage.html>

20. *UKOLN metadata resources—Dublin Core, list of projects*. Available from: <URL:http://www.ukoln.ac.uk/metadata/resources/dc.html>

21. *Europagate*. Available from: <URL:http://europagate.dtv.dk/

22. *UKOLN Experimental Z39.50 based demonstrators*. Available from: <URL:http://roads.ukoln.ac.uk/cgi-bin/egwcgi/egwirtcl/targets.egw>

23. Allen, J. and Mealling, M., *The Architecture of the Common Indexing Protocol (CIP)*. Internet-Draft, 9 June 1997. Available from: <URL: ftp://ds.internic.net/internet-drafts/draft-ietf-find-cip-arch-00.txt>

24. Joint Funding Councils' Libraries Review Group, *Report* [The Follett Report]. Bristol: Higher Education Funding Council for England, December 1993, Section 265.

25. Electronic Libraries Programme, *Project details*. Available from:<URL:http://www.ukoln.ac.uk/services/elib/projects/>

26. *ROADS*. Available from: <URL:http://www.ukoln.ac.uk/roads/>

27. Knight, J. and Hamilton, M., *Overview of the ROADS software*. LUT CS-TR 1010. Loughborough: Loughborough University of Technology, March 1996. Available from: <URL:http://www.roads.lut.ac.uk/Reports/arch/arch.html>

28. *NewsAgent for Libraries*. Available from: <URL:http://www.sbu.ac.uk/~litc/newsagent/>

29. *Harvest Web Indexing*. Available from: <URL:http://www.tardis.ed.ac.uk/harvest/>

30. *DESIRE*. Available from: <URL:http://www.nic.surfnet.nl/surfnet/projects/desire/desire.html>

31. *DESIRE WP3 Resource Discovery and Indexing*. Available from: <URL:http://www.ub2.lu.se/desire/>

32. Koch, T., Ardö, A., Brümmer, A. and Lundberg, S., *The building and maintenance of robot based Internet search services: a review of current indexing and data collection methods*. Draft D3.11 (version 3) for Work Package 3 of Telematics for Research project DESIRE, September 1996. Available from: <URL:http://www.ub2.lu.se/desire/radar/reports/D3.11/>

33. Lundberg, S., Ardö, A., Brümmer, A. and Koch, T., *The European Web Index: an Internet search service for the European higher education, research and development communities*. Deliverable 3.1 for Work Package 3 of Telematics for Research project DESIRE, 1996. Available from: <URL:http://www.nic.surfnet.nl/surfnet/projects/desire/deliver/WP3/D3-1.html>

34. *Specification for resource description methods*. Deliverable for Work Package 3 of Telematics for Research project DESIRE, February-May 1997. Available from: <URL:http://www.ukoln.ac.uk/metadata/DESIRE/specification.html>

35. *BIBLINK*. Available from: <URL:http://www.ukoln.ac.uk/metadata/BIBLINK/>

36. Heery, R., *Metadata formats*. Work Package 1 of Telematics for Libraries project BIBLINK (LB 4034), November 1996. Available from: <URL:http://www.ukoln.ac.uk/metadata/BIBLINK/wp1/d1.1/>

37. Høgås, H., van der Werf, T. and Powell, A., *Identification*. Work Package 2 of Telematics for Libraries project BIBLINK (LB 4034), May 1997. Available from: <URL:http://www.ukoln.ac.uk/metadata/BIBLINK/wp2/d2.1/>

38. *Nordic Metadata Project*. Available from: <URL:http://linnea.helsinki.fi/meta/>

39. *Arts and Humanities Data Service*. Available from: <URL:http://ahds.ac.uk/>

40. Dempsey, L., and Greenstein, D., *Proposal to identify shared metadata requirements*. 15 January 1997. Available from: <URL:http://www.kcl.ac.uk/projects/ahds/jobs/proposal.html>

41. Miller,P., *Resource Discovery Workshops: a guide to implementation and participation*. 23 May 1997. Available from: <URL:http://www.york.ac.uk/~apm9/focus01.html>

42. *MathN Broker*. Available from: <URL:http://www.mathematik.uni-osnabrueck.de/harvest/brokers/MathN/>

43. Plümer, J. and Schwänzl, R., *A mathematics preprint index: DC in an application*. [Paper for: 4th Dublin Core Metadata Workshop, Canberra, 3-5 March 1997]. Available from: <URL:http://www.dstc.edu.au/DC4/roland/>

44. *Scheme Definition: DMV MetaData for Mathematical Papers*, Version 1.2. Available from: <URL:http://www.mathematik.uni-osnabrueck.de/ak-technik/DMVPreprint-Core.html>

45. Dempsey, L., Distributed library and information systems: the significance of Z39.50. *Managing Information*, 1(6), June 1994, 41-43.

46. Turner, F., *An overview of the Z39.50 Information Retrieval standard*. IFLA Universal Dataflow and

Telecommunications Core Programme, Occasional Paper, 3, July 1995, rev. January 1997. Available from: <URL:http://www.nlc-bnc.ca/ifla/VI/5/op/udtop3.htm>

47. Raggett, D., Le Hors, A. and Jacobs, I.(eds.), *HTML 4.0 Specification*. W3C Working Draft. 18 July 1997. Available from: <URL:http://www.w3.org/TR/WD-html40-970708/>

48. *XML White Paper*. Microsoft Corporation, June 23,1997. Available from: <http://www.microsoft.com/standards/xml/xmlwhite.htm>

49. Resnick, P. and Miller, J., PICS: Internet access controls without censorship. *Communications of the ACM*, 39 (10), October 1996, 87-93

FURTHER READING

Metadata is one of those subjects that has a rapidly growing literature and is also an area which has regular changes of focus and emphasis. As can be seen by the references in this Briefing, a large amount of information on metadata topics is available on the Internet and specifically through the World Wide Web. For these reasons it may be useful to note the following Web sites devoted to keeping up-to-date with the subject:

- International Federation of Library Associations. *DIGITAL LIBRARIES: Metadata Resources*. Available from: <URL:http://www.nlc-bnc.ca/ifla/II/metadata.htm>
- UKOLN Metadata Group. *Metadata*. Available from: <URL:http://www.ukoln.ac.uk/metadata/>

Dempsey, L., 'Meta Detectors'. *Ariadne*, 3, May 1996, 6-7. Available from: <URL:http://www.ukoln.ac.uk/ariadne/issue3/metadata/>

Dempsey, L., 'ROADS to Desire: some UK and other European metadata and resource discovery projects'. *D-Lib Magazine*, July/August 1996. Available from: <URL:http://www.dlib.org/dlib/july96/07dempsey.html>

Dempsey, L., and Heery, R., 'Metadata: a current view of practice and issues'. *Journal of Documentation* (forthcoming).

Dempsey, L. and Weibel, S., 'The Warwick Metadata Workshop: a framework for the deployment of resource description'. *D-Lib Magazine*, July/August 1996. Available from: <URL:http://www.dlib.org/dlib/july96/07weibel.html>

Heery, R., 'Review of metadata formats'. *Program*, 30(4), October 1996, 345-373.

Lynch, C., 'Searching the Internet'. *Scientific American*, 276(3), March 1997, 44-48. Also available from: <URL:http://www.sciam.com/0397issue/0397lynch.html>

Wallace, D., 'Metadata and the archival management of electronic records: a review'. *Archivaria*, 36, Autumn 1993, 87-110.

Weibel, S., 'The World Wide Web and emerging Internet resource discovery standards for scholarly literature'. *Library Trends*, 43(4), Spring 1995, 627-644.

Weibel, S., 'Metadata: The Foundations of Resource Description'. *D-Lib Magazine*, July 1995. Available from: <URL:http://www.dlib.org/dlib/July95/07weibel.html>

ACKNOWLEDGEMENTS

UKOLN is funded by the Joint Information Systems Committee of the Higher Education Funding Councils and by the British Library Research and Innovation Centre, as well as by project funding from several sources.

The work carried out in this document is supported by the ROADS, DESIRE and BIBLINK projects.

The authors would like to thank Lorcan Dempsey for commenting on a draft version of this Briefing.